HPC OS Directions and Requirements

HPC User Forum, April 2008

John Hesterberg

Processor-rich environments

- Today's max SGI system sizes:
  - 1000s of cores per ccNUMA Single System Image (SSI)
    - •2048 -> 4096 cores
    - •limited by Linux scalability
  - 1000s of small SSI nodes per system
    - •14336 cores
    - •limited by cost
- Mutli-core is another layer of NUMA! In one SSI:
  - threads in a core
  - **cores in a cpu/socket**
  - sockets on a numa (or bus) node
  - numa nodes in a SSI
    - •latency/bandwidth between nodes
  - SSIs in a system

sgi INNOVATION FOR RESULTS

Processor-rich environments

- Job placement, scheduling continue to be critical, both within SSI and across SSIs
  - batch managers **need** to see topology
  - placement policies based on applications
- within a node:
  - processor/cache/node affinity
  - cpusets
  - isolated cpus
- across nodes:
  - synchronized scheduling

sgi INNOVATION FOR RESULTS

Resiliancy requirements

- all hardware is not created equal. :)
  ... but all hardware will fail
- tolerate as many failures as possible
  - memory in particular
  - diskless clusters eliminate one more failure component
  - this is an ongoing (and endless) effort
  - tight integration between software and hardware helps
- as systems scale, node failures become a certainty
  - application based checkpoint/restart?
  - OS supported checkpoint/restart?

The Power of Linux

- Disruptive OS development process
  - >1000 developers, 186 known companies (2.6.24)
    - https://www.linux-foundation.org/publications/linuxkerneldevelopment.php
  - That's just the kernel!!!
- Code reviews and signoffs
  - Hard and controversial changes get reviewed and done right
- Open Source Flexibility
  - Can be customized for HPC hardware
  - Can be on the cutting edge of research
- Productized and supported
- Same OS on laptop, desktop, clusters with 1000s of nodes, SSIs with 1000s of cores
- No vendor lock-in.

sgi INNOVATION FOR RESULTS

Virtualization

- Microkernel approaches (e.g. Xen)
  - hides hardware details :-(
  - how much memory do you want to waste?
  - how much cpu do you want to waste?
  - what about IO performance?
- KVM more interesting for some problems
  - First domain runs directly on hardware
  - Additional domains for compatibility.
- Rebooting
  - Can be better for clusters with dedicated nodes...compare boot times to job runtimes
  - No added runtime overhead
  - Provides clean image
  - Multiple boot images on disk or on server
  - Integrate into batch manager!

sgi INNOVATION FOR RESULTS

Green Computing

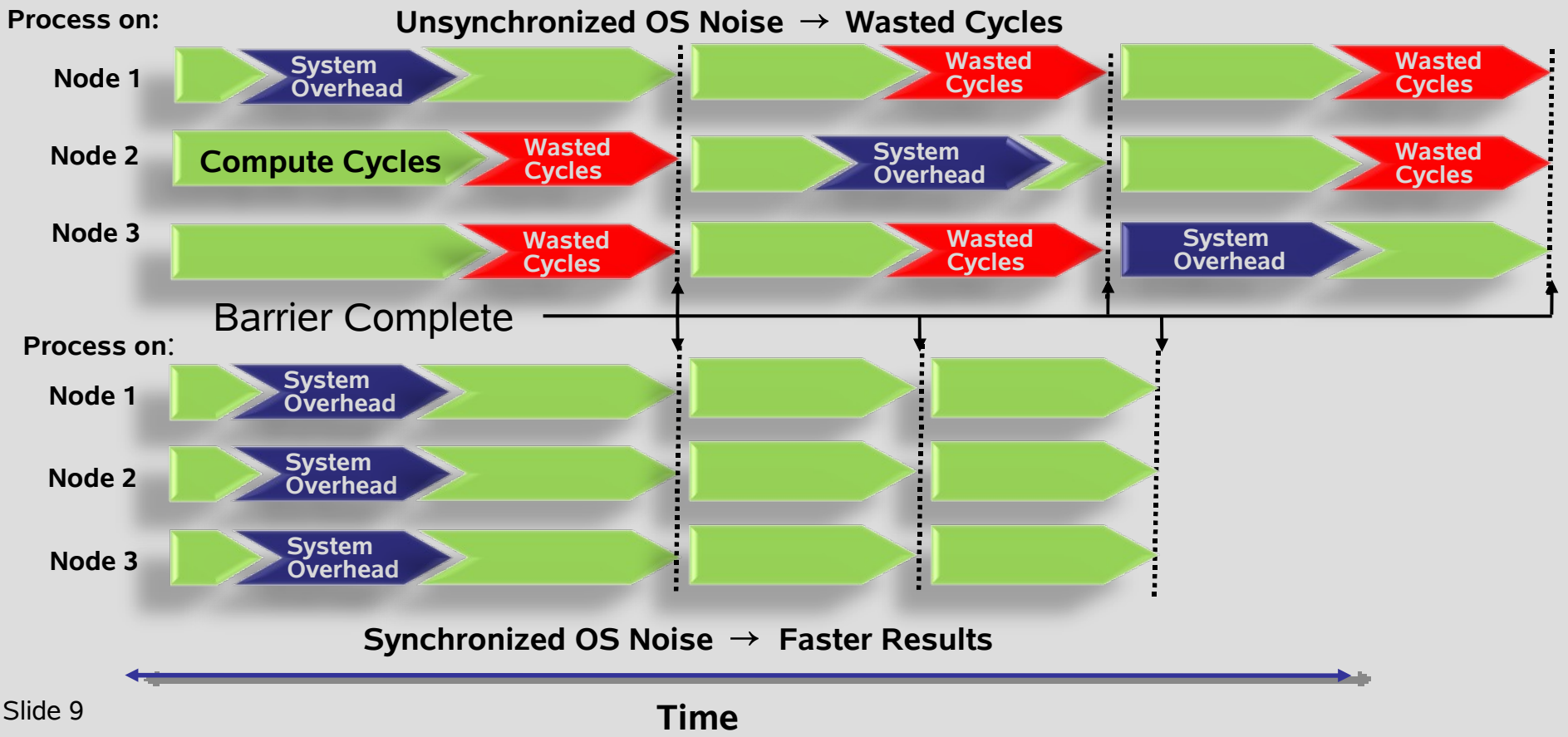
- hardware is helping and hurting...
  - laptop capabilities moving up
  - more and larger systems consuming more power
- HPC challenge: save power w/o sacrificing performance
  - tickless OS
  - cpu frequency management
  - management of idle cpus, nodes, and systems

sgi INNOVATION FOR RESULTS


sgi®

OS Noise Synchronization

- OS system noise: CPU cycles stolen from a user application by the OS to do periodic or asynchronous work (monitoring, daemons, garbage collection, etc).
- Management interface will allow users to select what gets synchronized
- Huge performance boost on larger scales systems



Servers: One Size Doesn't Fit All!

Workflow Characteristics

- Multi-Discipline
- Data-Intensive
- Mixed or Uncertain Workloads
- Interactivity
- Rapid development cycles

Workflow Characteristics

- Price-performance key
- Little data sharing
- Predictable Workloads
- Non-interactive
- Standard Modes
- Mature Apps

sgi INNOVATION FOR RESULTS

User Productivity Advantages of Large Global Shared Memory

- Freedom to **arbitrarily scale** problem size without decomposition or other rework
- Minimal penalty to fetch off-node data
- Ability to **freely exploit Open MP and/or MPI** in any combination or scale.
- **Simplified code development** and prototyping platform
- Freedom to experiment without the hindrance of cluster paradigms
- Unified parallel C translator in development
- Greatly **simplified load balancing**
- Simple to direct a task to any processor as all data is accessible

sgi INNOVATION FOR RESULTS